World Wide Web (2013) 16:421–447
DOI 10.1007/s11280-012-0177-1

Creation and growth of online social network

How do social networks evolve?

Katarzyna Musial · Marcin Budka · Krzysztof Juszczyszyn

Received: 7 January 2012 / Revised: 3 May 2012 /
Accepted: 25 June 2012 / Published online: 7 July 2012

Abstract Social networks are an example of complex systems consisting of nodes that can interact with each other and based on these activities the social relations are defined. The dynamics and evolution of social networks are very interesting but at the same time very challenging areas of research. In this paper the formation and growth of one of such structures extracted from data about human activities within online social networking system is investigated. Dynamics of both local and global characteristics are studied. Analysis of the dynamics of the network growth showed that it changes over time—from random process to power-law growth. The phase transition between those two is clearly visible. In general, node degree distribution can be described as the scale-free but it does not emerge straight from the beginning. Social networks are known to feature high clustering coefficient and friend-of-a-friend phenomenon. This research has revealed that in online social network, although the clustering coefficient grows over time, it is lower than expected. Also the friend-of-a-friend phenomenon is missing. On the other hand, the length of the shortest paths is small starting from the beginning of the network existence so the small-world phenomenon is present. The unique element of the presented study is that the data, from which the online social network was extracted, represents interactions between users from the beginning of the social networking site existence.

K. Musial (✉)
Department of Informatics, School of Mathematical and Natural Sciences,
King's College London, Strand Campus, WC2R 2LS London, UK
e-mail: katarzyna.musial@kcl.ac.uk

M. Budka
Smart Technology Research Centre, Bournemouth University,
BH12 5BB Poole, UK

K. Juszczyszyn
Faculty of Computer Science and Management,
Wroclaw University of Technology,
Wyb. Wyspianskiego 27, Wroclaw, Poland

The system, from which the data was obtained, enables users to interact using different communication channels and it gives additional opportunity to investigate multi-relational character of human relations.

Keywords online social network · complex system dynamics · network growth and evolution · dynamics of relationships of different types · local and global network characteristics

1 Introduction

The increasing complexity, dynamics and evolution of real-world networks combined with our constantly growing capabilities of gathering network data from natural and technology-based networked systems make network analysis one of the key challenges in the area of complex systems. The analysis of real-life complex networks is at the early stages and requires a lot of effort in both developing the tools and approaches to tackle them as well as understanding the nature and functioning of such networks.

Each complex networked system consists of large number of connected dynamic units whose behaviour is time-dependent, i.e. time factor cannot be neglected during analysis. The structure of complex networks is irregular and constantly evolving. The organization of these networks typically implies a skewed distribution of relations with many hubs, strong heterogeneity and high clustering as well as non-trivial temporal evolution.

Existing methods and models that have been developed to help us understand the changes occurring in complex networks are only partially useful for modelling of social systems. One of the elements that has not been investigated is how the dynamics of the evolution changes. The intuition suggests that social networks evolve at different pace at different development stages but there is no reported research in this area that confirms or rejects this statement.

One of the main goals of this study is to investigate the growth of real-world network from its beginning and understand the dynamics of its evolution. Moreover, the system under consideration enables the users to participate in more than one activity, i.e. people can contact each other directly by adding others to the contact list or post their views and opinions at the forum. This gives another view on social relationships as people can use different communication channels to exchange information, and this is an important factor that should also be taken into consideration during the analysis.

Over the last few decades, the online social networks have become a very important element of World Wide Web and they are a source of large amount of information that can be used to shape the future Wisdom Web of Things [27]. Online social networks mining is one of the Web Intelligence tools [28]. Analysis of people behaviours and interactions in the online world as well as how these behaviours change over time is a crucial element when it comes to create and compose personalised services in the Wisdom Web of Things.

2 Related work

The main area of this study is the growth and its dynamics of complex social networks where more than one type of relationship can exist. In this section first the concept of Complex Social Multinetwork (CSM) is briefly presented. After that different methods that enable to investigate the dynamics of social networks are discussed in order to introduce the reader into the world of network evolution.

2.1 Complex Social Multinetwork

Each Complex Networked System (CNS) consists of multiple interacting components whose global behaviour cannot be simply inferred from the behaviour of the individual parts [2, 11]. One of the examples of CNS are social networks that consist of nodes (social entities: humans or groups of people) and relationships (edges) linking pairs of nodes [23]. The whole research field of social network analysis has been developed over the years and its goal is to investigate social networks [6, 8, 9, 23]. Although social networks are one of the categories of complex networks, Newman and Park claim that the they differ from most other types of networks [20] because: (i) they show high clustering and/or network transitivity, and (ii) they show positive correlations also called assortative mixing between the degrees of adjacent vertices [19]. A relatively recent trend in the complex networked systems research is the exploration of social media that allow users to interact and collaborate with each other in many different ways, both directly and indirectly [3, 21, 26]. For instance, a social networking service allows to publish photos, comment and tag them, mark them as favourite, add other user to contact list, join user groups, comment profiles or photos, categorize photos, post in topics, etc. [12, 22]. The extracted from data social networks can be usually both multirelational and multimodal. The former ones are networks that consist of more than one type of relationship while the latter have more than one type of node. Different relations can emerge from different communication channels, i.e. based on each communication channel separate relation is created. Different nodes can be extracted from different systems, e.g. a set of email users and a set of blog users. These systems are known as Complex Social Multinetworks and their concept is presented in Figure 1, where the *id* denotes a user identifier in a given network (e.g. an email address in an email network). Each user can have more than one identifier, e.g. user x and user y have two *ids* as they participate in two different networks (e.g. networks created based on communication using email and instant messenger). The set of *ids* of a given user is called virtual identifier (*vid*). $R_1, \cdots, R_7$ denote different types of relationships that exist between users. $R_1, \cdots, R_3$ are the relation types that can be distinguished in the first multirelational social network $MRSN_1$ and those of type $R_4, \cdots, R_7$ exist in $MRSN_2$. For more information about this concept please see [17].

The network that has been analysed in this study features different types of relations but the set of nodes is the same for all relation types, i.e. it is a multirelational network.

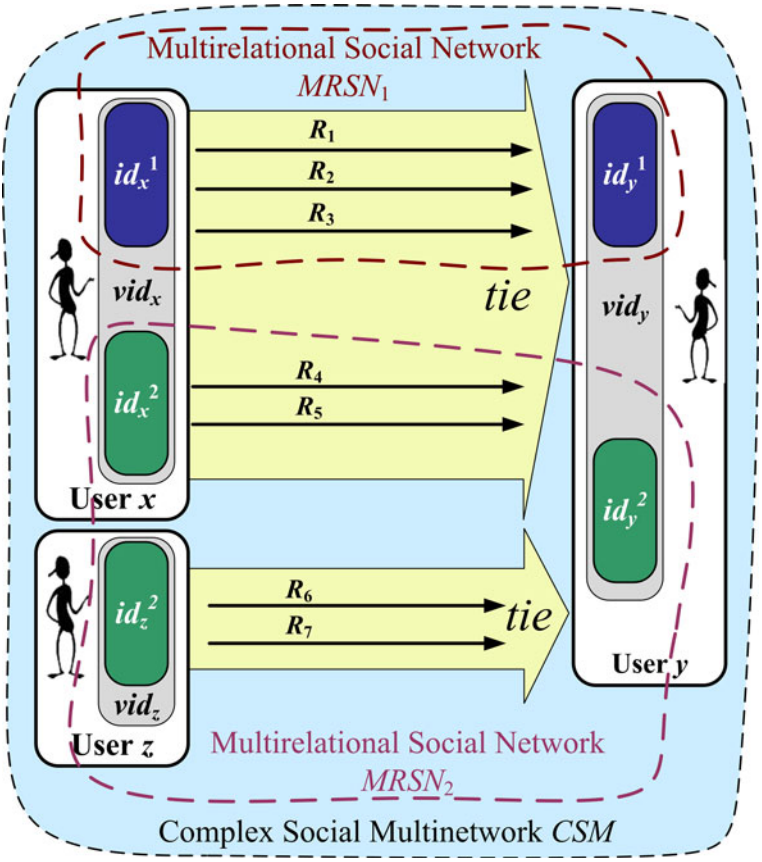

Figure 1 Concept of Complex Social Multinetwork.

2.2 Dynamics of social networks

In the last few years the problem of investigating dynamics of social networks has become an important research challenge. Most of the approaches that try to address the complex network growth take into consideration some global characteristics of the networks and develop models that reproduce these characteristics, e.g. node degree distribution [1], clustering coefficient [24] or network diameter [4]. There are some approaches that aim at developing specific models for online social networks and take into consideration some information characteristic to such networks [5, 7, 14–16].

In [14], based on the analysis of real-world networks such as Flickr and Yahoo 360!, users have been divided into three different types: passive, linkers and inviters. Authors define system that follows specific set of rules used to describe the evolution of the social network. The method that describes the network growth can be defined as the set of steps: (i) at each time step, a node arrives, and one of the statuses: passive, linker, or inviter is randomly assigned to it; and (ii) during the same time

step, x edges arrive and the source of each of the edges is chosen at random from the existing inviters and linkers in the network using preferential attachment. Depending on the chosen type of source node (inviter or linker) different actions is performed.

In [7] researchers focus on discovering patterns of interactions between users and their evolution over time. Authors propose to create a graph that represents a social network but additional information is the time-stamp added to each relation when it appears in the network for the first time. Similarly to the previous presented study, also this one assumes that the users and the relations between them can only be added to the system and will never disappear.

Another framework for network growth was developed in [15] where authors studied four online social networks: Flickr, Delicious, Answers and LinkedIn. They proposed to apply the maximum-likelihood estimation principle to compare a family of parameterised models in terms of their probability of generating the observed data, and as a result to select the model that reflects data in the best possible way. For every edge arriving to the network the probability that it will connect two given nodes under some model is assessed. The product of these probabilities over all edges gives the likelihood of the model and the model with the highest probability is perceived as the best one.

Yet another set of approaches that take into consideration the fact that links can disappear from the network are those which exploit the dynamic centrality phenomena [5, 10]. In [5] authors have detected a dramatic time dependance in network centrality and the role of nodes, something that is not apparent from the static analysis of node connectivity and network topology. Authors found that the daily networks were scale-free but the well-connected nodes from these networks changed from day to day.

Although all the methods described above model the growth of the networks, none of them investigates the dynamics of this growth. This study analyses the evolution of a network starting from its creation and aims to measure the dynamics of the network growth.

3 Methodology

The goal of this research is to analyse how the network evolves from the very beginning, from the first relation that has appeared within the network. The research methodology employed in this work is outlined in Figure 2. The building blocks of the methodology are: (i) data preparation and network extraction process, (ii) network characteristics that will be investigated, and (iii) plan of the experiments together with their further analysis and evaluation.

3.1 Data description

The system from which the data has been obtained is a social networking site for people who are building or improving their houses (*extradom.pl*[1]). Members can

[1] Queries regarding the access to the dataset should be sent to katarzyna.musial@kcl.ac.uk.

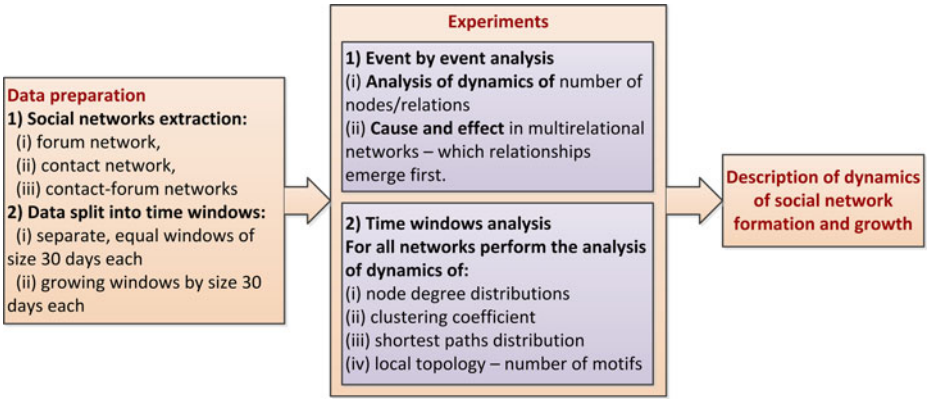

Figure 2 Research methodology.

share their experiences, discuss topics related to different projects of houses, share photos, comment others projects and participate in the forum. The data that has been used were obtained in the *.sql format. Information about contacts includes such information as identifiers of users in a relationship and the time stamp of the connection creation (int user_id; int contact_user_id; datetime creation_date). Data regarding forum contains identifiers of each topic, post and its author as well as time stamp of the post creation (int post_id; int topic_id; int post_author; datetime creation_date). The database schema of the analysed data is presented in Figure 3.

3.2 Network extraction

The first step is to extract the social networks to be analysed from the available data. The subset of data that was analysed includes information how people participate in different posts in the forum and who is connected to whom in terms of the contact

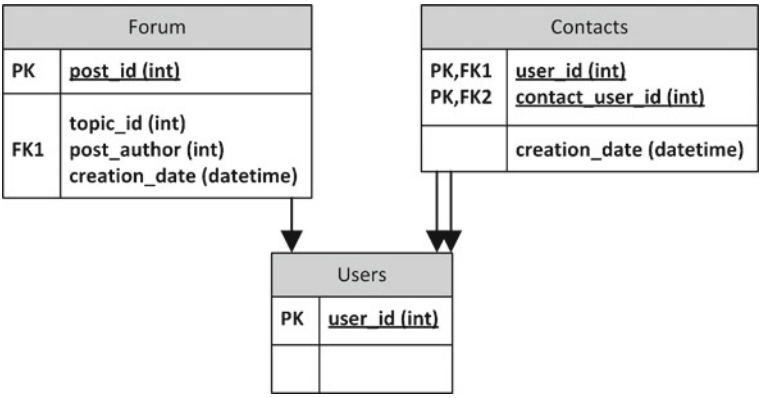

Figure 3 Used database schema.

list. During this study three networks have been created: (i) network based on the posts within the topics at the system forum, (ii) network based on the contact list in the system, and (iii) network created by merging contact and forum networks. The relationship exists in the contact network if there is information about their contact in the contact list that each user creates (similar to the list of friends on Facebook). The edge in the forum network is created if people post in the same topic within the forum. In the merged network the relationship between two nodes exists only if it is present in both contact and forum networks. The network extraction process is presented in Figure 4.

3.3 Network characteristics

The characteristics that were analysed in this study are: number of nodes and edges, node degree distribution, network density, shortest paths, clustering coefficient and triads.

Network density (D) is expressed as ratio of the number of connections in a given graph to a number of possible connections within this graph—$D = \frac{2 \cdot E}{V \cdot (V-1)}$, where E—no. of relations; V—no. of nodes.

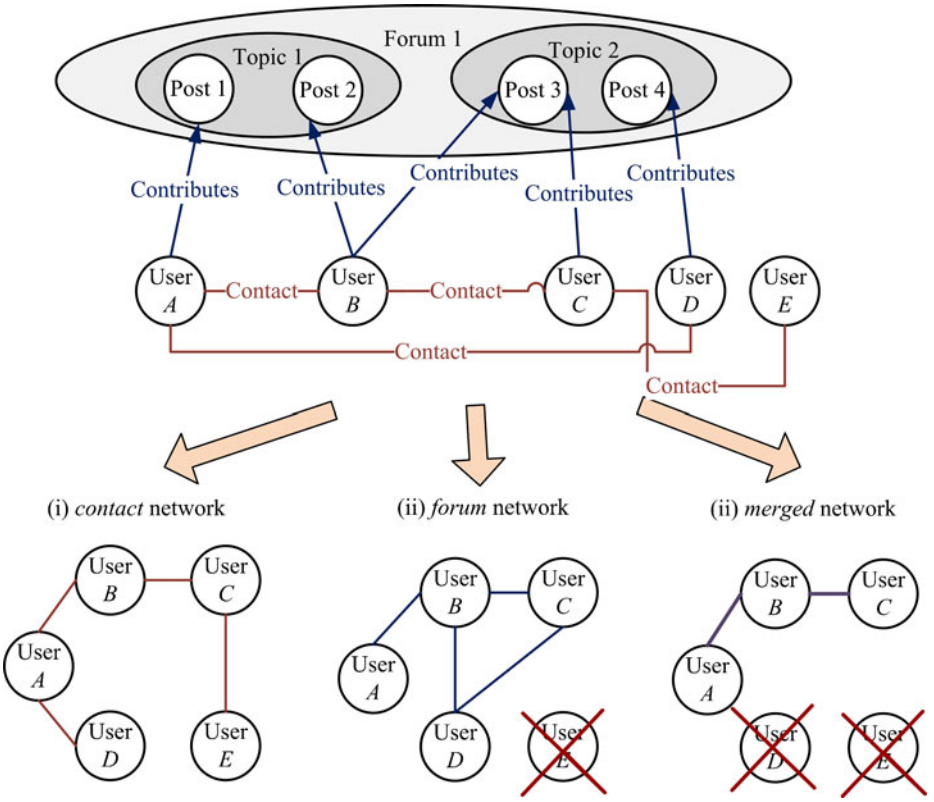

Figure 4 Networks extraction process.

The concept of the Shortest Paths (*SP*) is used to define whether the network features the small-world phenomenon i.e. whether it is easy or not to reach any other node in the network. The shortest path between two nodes is defined in this study as the minimum number of connections that separates these two nodes.

Clustering Coefficient (*CC*) is the next characteristic that is investigated in this paper. Suppose a vertex v has neighbours $\mathcal{N}(v)$, with $|\mathcal{N}(v)| = k_v$. At most $k_v(k_v - 1)/2$ edges can exist between them (this occurs when v is part of a k_v-clique). The clustering coefficient [25] of the vertex, C_v, is defined as the fraction of these edges that actually exist.

Finally, the dynamics of the local structures—motifs of size 3 (a.k.a. triads) will be performed. Triads are subgraphs that consist of three nodes. As the networks are undirected, only 8 different triads can be distinguished. All triad structures are presented in the Figure 5. Note that the sets of triads (2, 3, 5) and (4, 6, 7) are topologically equivalent. However, in the analysis the decision was made to distinguish them as the nodes in the social networks are labelled.

3.4 Experimental set-up

Experiments were performed in two blocks: (i) event-by-event analysis and (ii) time windows analysis. The goal of the first set of experiments was to investigate how the number of nodes, interactions and in consequence number of relationships change over time. Also the merged contact-forum network was investigated in order to find out which relationships (in contact list or at forum) are created first.

In the second set of experiments further dynamics of extracted social networks were analysed. In order to do that the whole dataset has been divided into time windows using two different techniques. The first one was to divide available data into separate, disjoint windows of size 30 days each. The second approach was to use a growing window, where the initial window covered 30 days and each consecutive has been created by taking the previous one and appending data from the next 30 days. For all the networks node degree distribution, clustering coefficient, shortest paths distribution, and local topology features (triad count) were evaluated in each

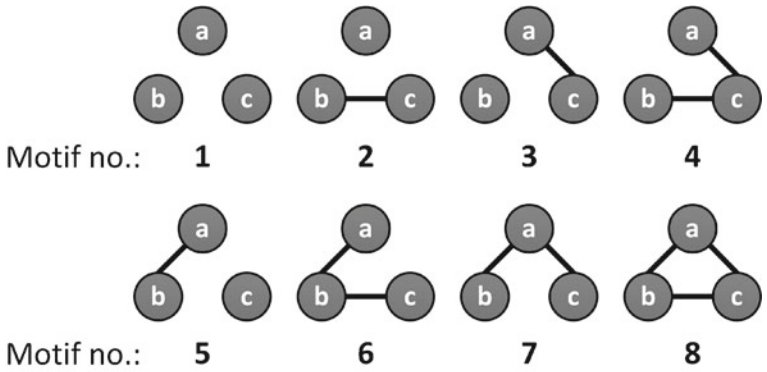

Figure 5 Analysed triads and their numbers used in the further part of the study.

of the extracted windows. Note that 16 time windows were created and this covers first 480 days of network existence. The remaining data that covers 26 last days in the dataset was not analysed as it was not possible to create based on it 30 day time window. This data has been neglected to ensure both the consistency of the approach and the outcomes of the analysis.

Please note that, the size of time window for splitting the data has to be carefully selected. Too large window will cause that the interesting patterns can be not visible and too small window can bring too much noise to the analyses. The previous study on this dataset revealed that in the 90-day time window an interesting seasonality can be observed [13]. This was caused by the fact that people tend to build houses in the spring and summer period whereas during winter their activity in this direction is much smaller.

In this study, in the case of separate time windows, the dataset has been split into windows of size 30 because the preliminary study showed that smaller windows result in networks where no interesting patterns in terms of number of nodes and edges could be found. Please note however, that this paper is focused on growth of network so most of the experiments were performed on growing windows where next window contains the previous window. In such situation the size of the growth influences the granularity of information. In this case growth smaller than 30 days did not provide additional information and larger one caused that although the trends in network growth were visible, they had discrete character what makes them hard to analyse.

After performing all experiments, the results are analysed in terms of network dynamics and its character. Changes in local and global networks' characteristics enable to assess the changes in network dynamics.

4 Experiments analysis and discussion

The experiments and their outcomes are described in an order that is presented in Figure 2. In Section 4.1 the basic networks characteristics, such as number of nodes/edges, networks density are presented. After that the event by event analysis of extracted networks is discussed (Section 4.2). This includes changes in number of nodes and edges over time as well as the origin of the relationships. In Section 4.3 the time window analysis is performed where data has been divided into separate and growing time windows. The characteristics investigated in this part are: node degree distribution, shortest paths, clustering coefficient and finally the triads evolution.

4.1 Basic networks characteristics

The first stage of the experiments was to investigate the basic characteristics of networks extracted from the gathered data (see Table 1). Data from the system comes from the period between 21/08/2008 and 08/01/2010, which covers the first 16 months of its existence. The number of users of the system during this period is 103,716.

The number of nodes in the contact network is much bigger than in the forum network. In the contact list 102,928 users can be identified. It means that the network is very sparse as there is only 112,363 connections and the full graph for this number of nodes would have more than $6 \cdot 10^9$ connections. Similar situation appears within

Table 1 Characteristics of extracted contact and forum networks.

Characteristic	Contact network	Forum network
Time period	21/08/2008, 15:13:54–08/01/2010, 15:12:19	21/08/2008, 16:12:18–08/01/2010, 15:18:07
Number of nodes	102,928	3,535
Number of relationships	112,363	8,110
Number of interactions	112,363	23,060
Network density $D = \frac{2 \cdot E}{V \cdot (V-1)}$ E—no. of relations; V—no. of nodes	$1.78 \cdot 10^{-5}$	$2.47 \cdot 10^{-4}$

the forum network where the number of relations is 8,110 and there are more than $3 \cdot 10^7$ possible relations. Both networks are sparse—the network density D is $1.78 \cdot 10^{-5}$ for the contact network and $2.47 \cdot 10^{-4}$ for the forum network.

The number of interactions and relations for the contact network is exactly the same as one adds somebody to the contact list only once and there are no other possible interactions between people within the contact list. The situation is different when it comes to the forum network. Here the number of interactions is bigger than the number of relations as the interaction is defined as posting within the same topic and two users can post together multiple times within different topics.

4.2 Event by event analysis

4.2.1 Number of nodes in time

The changes of the number of nodes in time are presented in Figure 6. Both charts show how the number of users in the forum and contact networks grows in time, but chart (a) presents the changes in log–log scale and (b) in linear scale.

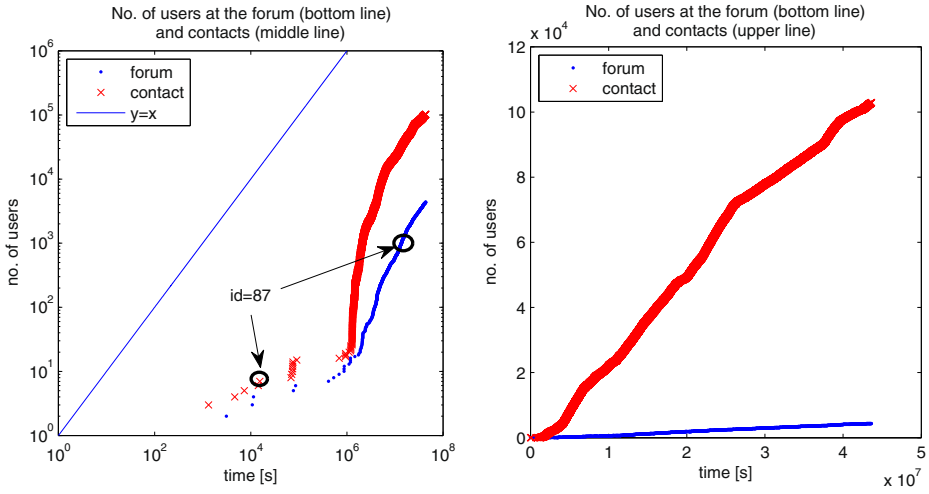

Figure 6 Growth of number of users in time for forum and contact list networks; **a** log–log scale, **b** linear scale. Time is expressed in seconds [s].

The contact network for the first 15 days grew by 50 users whereas during the first 23 days of network existence only 13 people posted at the forum.

In Figure 6 at the left chart the thin line shows the linear function $y = x$. It can be noticed that during the first period (up to day 15 in the contact network and up to day 23 in the forum network) the number of nodes grows much slower than the linear function, but after that the growth is faster than linear. Moreover, the number of nodes in contact list grows much faster than the number of users in the forum network. This can be caused by the fact that the main goal of people signing to the network is to seek advise regarding their house projects. Most of them add user with id 87 (marked in Figure 6), who is a consultant, to their contact list and this causes the boost in the number of users in the contact list network.

In addition, although people may read the forum, they tend not to participate, i.e. they do not post or comment the forum entries. It shows the nature of the network where people seek advise but do not intensively share their experiences. User 87 seems to highly influence the evolution of the network. Once s/he joined the network the number of contacts in the network increases. The fact that s/he has only one post on the forum may be the reason why it does not grow as fast as the contact network.

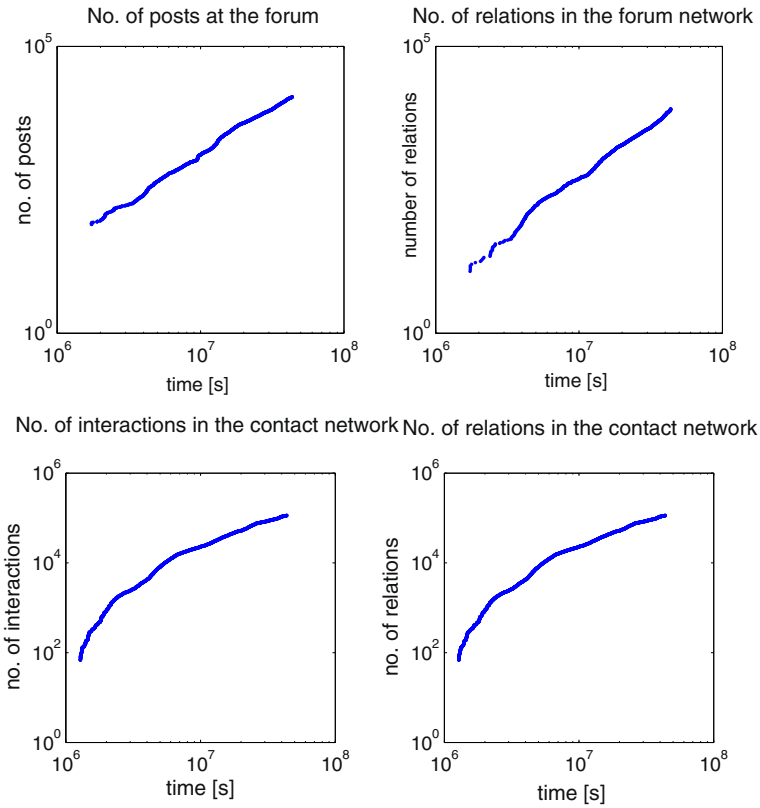

Figure 7 Changes of number of interactions/relations in time for forum and contact list networks. Time is expressed in seconds [s].

The communication on the forum is limited to occasional comments within a group of people who previously added themselves to each other contact list. Nevertheless, based on the analysed data user with id 87 seems to trigger the evolution of both contact and forum networks. The information about the professional consultant, who joined the systems, spread across the network and also in the real world. In consequence more people have started joining the network and the phase transition in the number of nodes occurred.

4.2.2 Dynamics of the number of relations

The second measure of network dynamics are the changes in the number of relations. There is 224,726 connections in the social network based on the contact list and 8,110 relations in the forum network. Note that these relations are undirected. The changes in the number of interactions (number of posts) and relationships in the forum network as well as in the number of both interactions and relations in contact network are presented in Figure 7. Note that number of interactions and relations for the contact network is really the same thing because the only interaction that can be recorded at the contact list is when one user adds another one to the contact list.

The situation with the number of relations is very similar to the one with the changes of the nodes number. At the beginning the growth is slow—only 79 interactions up to the 20th day of forum network life and 68 relations up to the 15th day of contact network existence. After that the number of relations grows faster. In the forum network this growth follows the power law with the exponent equal to 1.57 (i.e. it is super-linear) and linear growth can be observed for the contact network (Figure 8). This difference can be easily explained. When a user posts within a specific topic, the interaction is counted every time this user posts something. In the case of contact list one person adds another to the contact list only once. Also for the forum network, there is around three times more interactions in comparison to number of relations and it shows that there is no much communication within a single relationship. It further confirms the fact that people post on the forum when

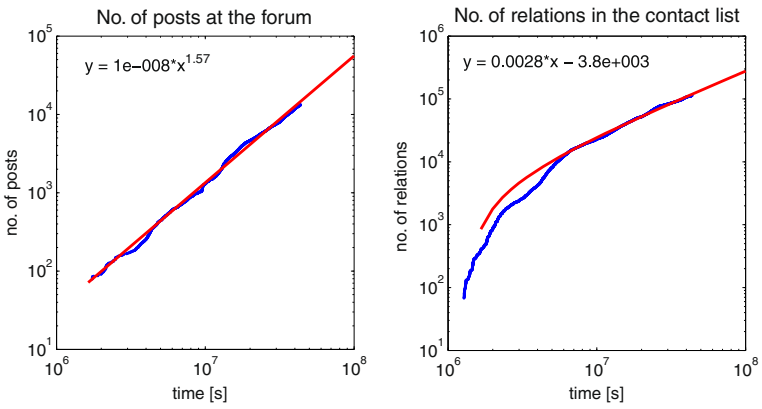

Figure 8 Changes of number of interactions in time for forum (after day 20th) and contact list (after day 15th) networks. Time is expressed in seconds [s].

they seek advice and when they get (or not) the answer they do not further sustain the relationship. This shows that the forum network has rather interest- than social-based character. This was also confirmed in the study on this dataset in [13].

An interesting phenomenon can be observed in the case of contact list as user with id 87 is involved in 102,800 of the relationships extracted in the contact network but he has only two relationships when it comes to the forum network. This person is a consultant who advises individual people but does not participate in general discussions at the forum.

4.2.3 Do relationships at forum emerge after relationships in contact list?

Another interesting feature is to investigate which relationships, these in contact list or these in the forum, emerge first. To analyse this the forum and contact networks were merged.

The number of relationships from the merged network, where a link between nodes in both forum and contact network was present, equals to 1,158. It constitutes only 0.52 % of all relations from the contact list and 14.29 % of all connections at the forum. It means that in most cases people maintain only one type of relationship. The experiments revealed that all relations from the merged network have been first created in the contact list and then the common activities in the forum have followed. If people interact using both types of relations then these connections origin from contact lists. This shows that relationships that are created first on forum never evolve into the direct relationship in the contact network and it seems to confirm the fact that the activities within the system are interest rather than social based.

4.3 Time windows analysis

4.3.1 Characteristics of networks in time windows

The general statistics regarding the number of nodes and distinct relations are given in Table 2 (separate windows) and Table 3 (growing windows).

Splitting data into separate time windows shows the level of user activity in specific time periods. In Figure 9 the number of nodes and edges in all networks for each time window is shown. For contact-forum network the number of nodes and edges grow from one window to another. Similar trend is visible in the case of the forum network. However, in this case between windows 6 and 12 we observe a “stability period”, where not a lot has changed in terms of the number of nodes and edges. Although the number of nodes/edges remains close to constant, it should be emphasized that the set of people participating in the forum in each of these time frames is significantly different. This leads us to the conclusion that even if the global features such as the number of nodes or edges are stable, it does not entitle us to say that the network does not change. The number of nodes and relations for the growing windows feature very similar pattern as shown in Figure 6.

The interesting conclusion from Figure 9 and based on the previous research on this dataset in [13] is that people tend to come to the network just for a specific period of time. In [13] it was shown that people tend to be active only within one time window (90 days).

Table 2 Characteristics of the separate 30 days time windows.

Window no.	No. of nodes contacts	No. of relations contacts	No. of nodes forum	No. of relations forum	No. of nodes merged	No. of relations merged
1	1,818	1,890	28	32	11	14
2	6,740	6,843	145	146	20	14
3	8,794	8,968	213	179	27	18
4	5,759	5,842	228	193	24	15
5	7,107	7,136	322	270	52	35
6	8,667	8,737	491	479	79	59
7	8,109	8,175	530	483	85	66
8	6,260	6,271	462	388	76	45
9	9,956	10,071	543	465	83	65
10	10,425	10,479	532	460	108	83
11	4,961	5,002	564	529	118	90
12	4,661	4,611	561	489	112	82
13	5,108	5,125	752	742	147	118
14	5,241	5,290	737	788	140	105
15	7,558	7,787	904	987	175	131
16	5,818	6,070	1,135	1,343	150	179

4.3.2 Degree distribution

The next characteristic that has been examined is the degree distribution. Considering the whole network, in the contact network there is one node with degree 102,800. The rest of the degrees are from the range [1; 244]. There are 98,010 nodes with degree 1 and it constitutes 95 % of all nodes. For more details please refer to Figure 10b.

Table 3 Characteristics of the time windows growing by 30 days.

Window no.	No. of nodes contacts	No. of relations contacts	No. of nodes forum	No. of relations forum	No. of nodes merged	No. of relations merged
1	1,818	1,890	28	32	11	14
2	8,541	8,733	158	175	27	25
3	17,183	17,701	307	340	42	37
4	22,745	23,543	450	519	57	48
5	29,650	30,679	640	766	90	77
6	37,980	39,416	932	1,217	129	115
7	45,626	47,591	1,217	1,667	173	160
8	51,533	53,862	1,453	2,016	205	185
9	61,084	63,933	1,720	2,441	242	229
10	70,955	74,412	1,966	2,859	285	279
11	75,310	79,414	2,205	3,329	332	329
12	79,400	84,025	2,400	3,747	375	368
13	83,837	89,150	2,662	4,394	414	413
14	88,247	94,440	2,880	5,074	459	456
15	94,949	102,227	3,118	5,914	512	511
16	99,740	204,910	3,374	7,064	553	557

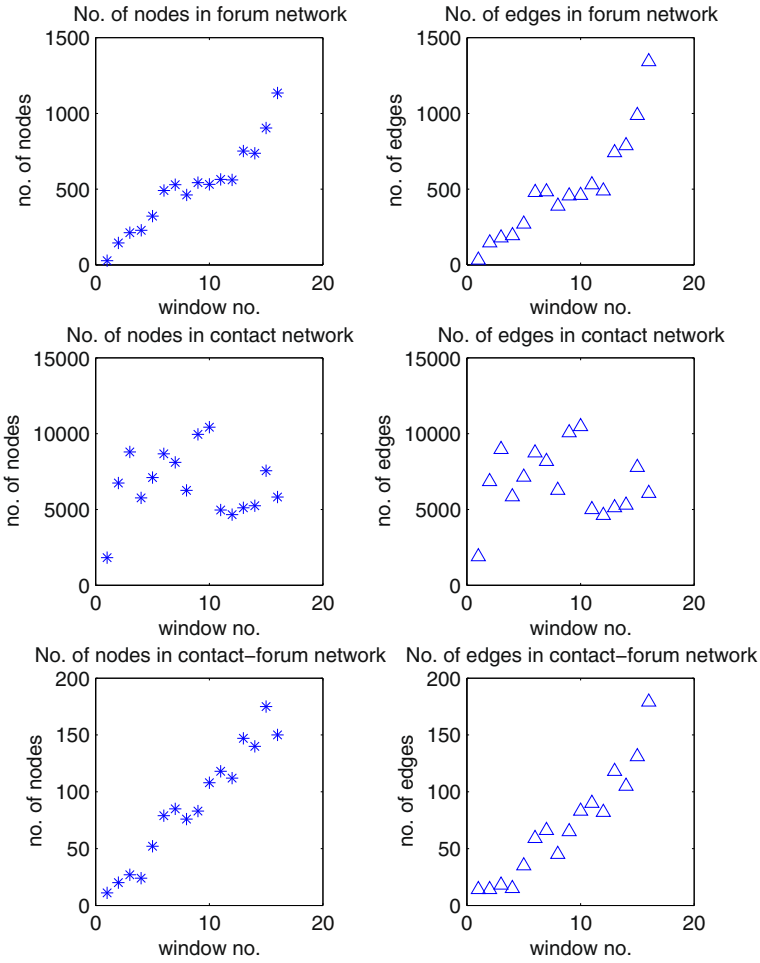

Figure 9 Dynamics of number of nodes and edges in windows of size 30 days.

In the forum network the node degree distribution is not as diverse as in the case of contact network. The highest node degree is 419 and there is one user with such high degree. The rest of the users have number of relationships between 1 and 291. Note that only 10 users have node degree above 100 (Figure 10a). Figure 10c presents one more node degree distribution where forum and contact networks were merged in a way that only relations which exist in both networks (forum and contact) were considered. In this network, the smallest node degree is 3 and its probability is 0.37. On the other hand the maximum degree is 261 and only one node has this number of connections.

Note that all the distributions are the long tail distributions where only a few nodes are highly connected and there is a lot of weakly connected individuals.

In order to investigate the dynamics of the node degree distribution of the networks, the growing windows were analysed (see Section 3 for more details). It

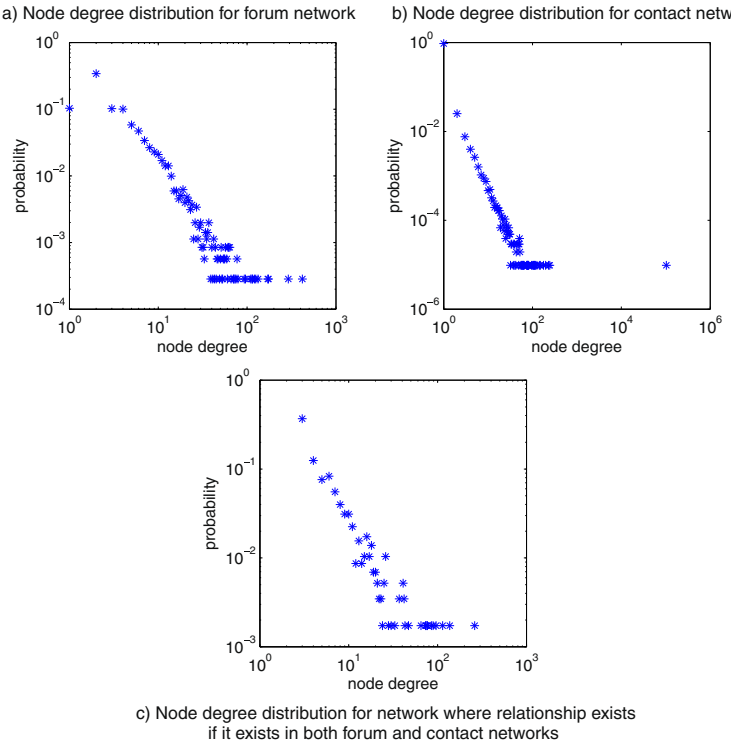

Figure 10 Node degree distribution for all extracted networks network for entire analysed period.

can be noticed that regardless the type of relationship the more nodes are added to the network, the more scale free distribution is visible (see Figures 11, 12 and 13).

The number of nodes with degree 1 is especially high for the contact network and it constitutes more than 95 % of all nodes for all windows. In the forum network the percentage of nodes with degree one decreases in time. In Window 1 it is 60 %, in Window 5 it is 52 %, in Window 10—48 %, in Window 15—30 %. Finally in Window 16 there is more nodes with degree 2 (31 %) than degree 1 (23 %). It shows that the network is getting more and more denser in time (this is also confirmed in the Section 4.3.3 where distribution of shortest paths are considered. In the merged network the percentage of nodes with degree one is more than 62 % and less than 71 % in all windows but it does not feature any interesting pattern.

4.3.3 Shortest paths

Analysis of the shortest paths allows to determine whether the networks feature the small-world phenomenon that is typical for social networks. In other words, this part of experiment examines the length of paths between the users. If the paths are short, then it is possible to reach other network members in just a few steps. The study also investigates the evolution of the probability distribution of shortest path length in time. The results are presented in Figures 14, 15 and 16 for contact, forum, and merged networks respectively.

The network based on contacts does not change over time in terms of shortest path length distribution. For each growing time window the maximum length equals to 5 and on average the shortest path has length 2. For all time windows except the first one, over 99 % of shortest paths have length 2. In the case of window 1, path of length 2 constitutes 97 % of all shortest paths. This is quite an interesting observation as the contact network is very sparse—its density equals to $1.78 \cdot 10^{-5}$ (see Section 4.1). It shows, that although basing on the global characteristic such as density one could say that the network is not well connected, further investigation into shortest paths revealed that traversal through this network can be done very effectively, as the shortest path between two users equals 2 on average and its maximum value is 5. This effect is caused by the user with id 87 who is in the relation with almost everybody in the network. As a result we observe a star structure with very small

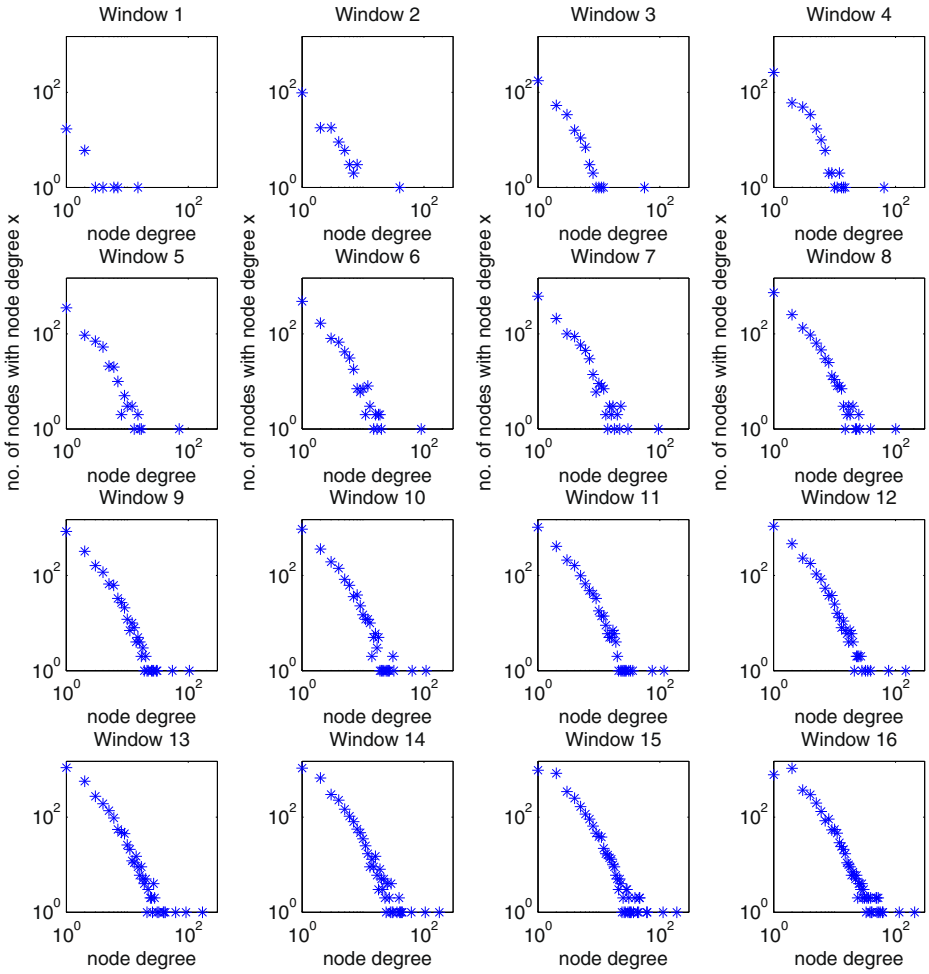

Figure 11 Degree distribution for forum networks for growing windows.

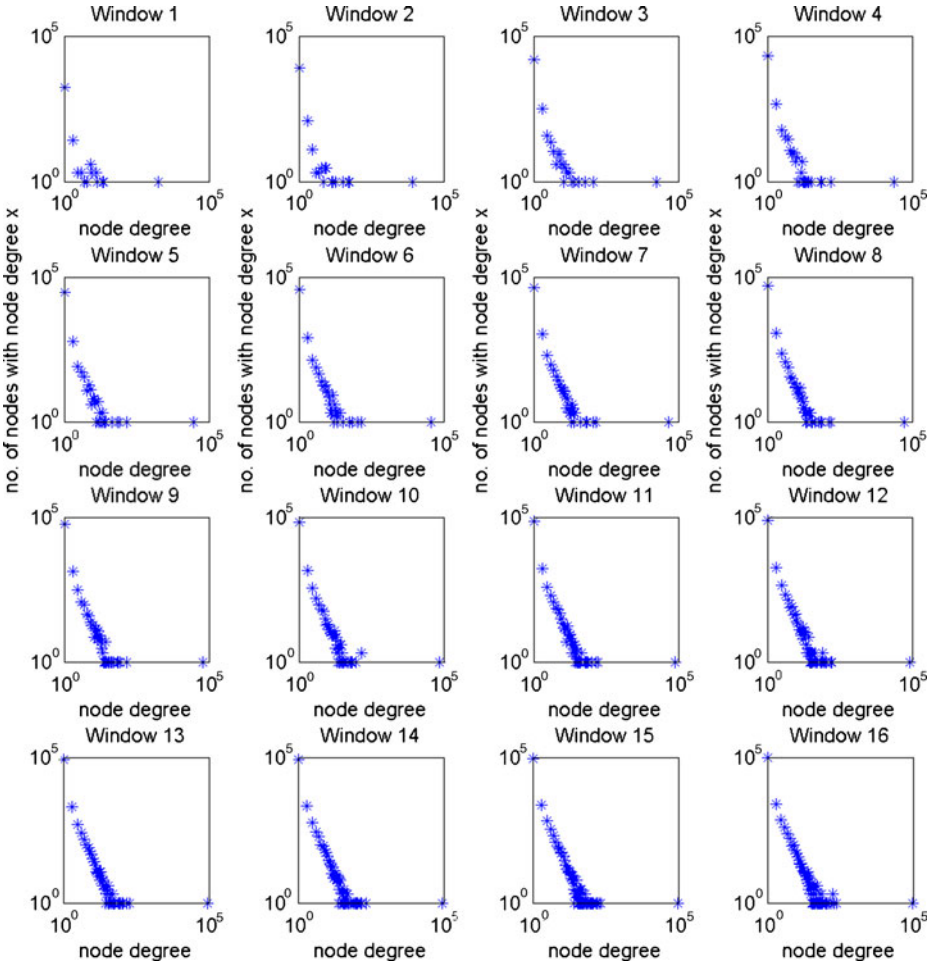

Figure 12 Degree distribution for contact networks for growing windows.

length of shortest path and very low clustering coefficient (see Section 4.3.4). This network can be described as degenerated scale-free network with only one hub.

The forum network exhibits much more variety in terms of shortest paths than the contact network. Here, the maximum length of the shortest path, a.k.a. diameter of the network, equals 21 and can be only observed in Window 8 with probability $1.461 \cdot 10^{-6}$. In the first time window shortest path of length 3 has the highest probability—0.41. In Window 2, length 3 is still the most probable (0.26) but in Windows 3 and 4 the path of length 4 (0.26 and 0.21 respectively). For the remaining time windows length 5 is the most probable (0.21). An interesting phenomenon can be observed for path of length 6 as its probability is only a little bit smaller than for the path of length 5. Starting from Window 8 the difference between probabilities for path length 5 and 6 decreases. This process reverses after Window 12, when the differences start to increase. This is the point in the evolution of the network where new relations are still created but the number of nodes does not increase as rapidly

as in the previous windows (8–12). See Table 3 where growth of number of nodes and relations is presented for each time window.

For the last, merged network the distribution of the shortest path length and its evolution is presented in Figure 16. As the number of nodes and relations grow in consecutive time windows the shortest paths get longer as well. In Window 1 the maximum length of shortest path is 4 and for Window 16 it equals 9. For windows 1–the shortest path with the highest probability (over 0.4 in each of the windows) has length 2. For window 5 length 3 dominates in the network with probability 0.26. For the rest of the time frames (6–16) length 4 of the shortest path has the highest probability—over 0.24 for each window.

Note that, for all networks the probability distribution of shortest path length has a shape of the normal distribution. Moreover, the these distributions peak at relatively

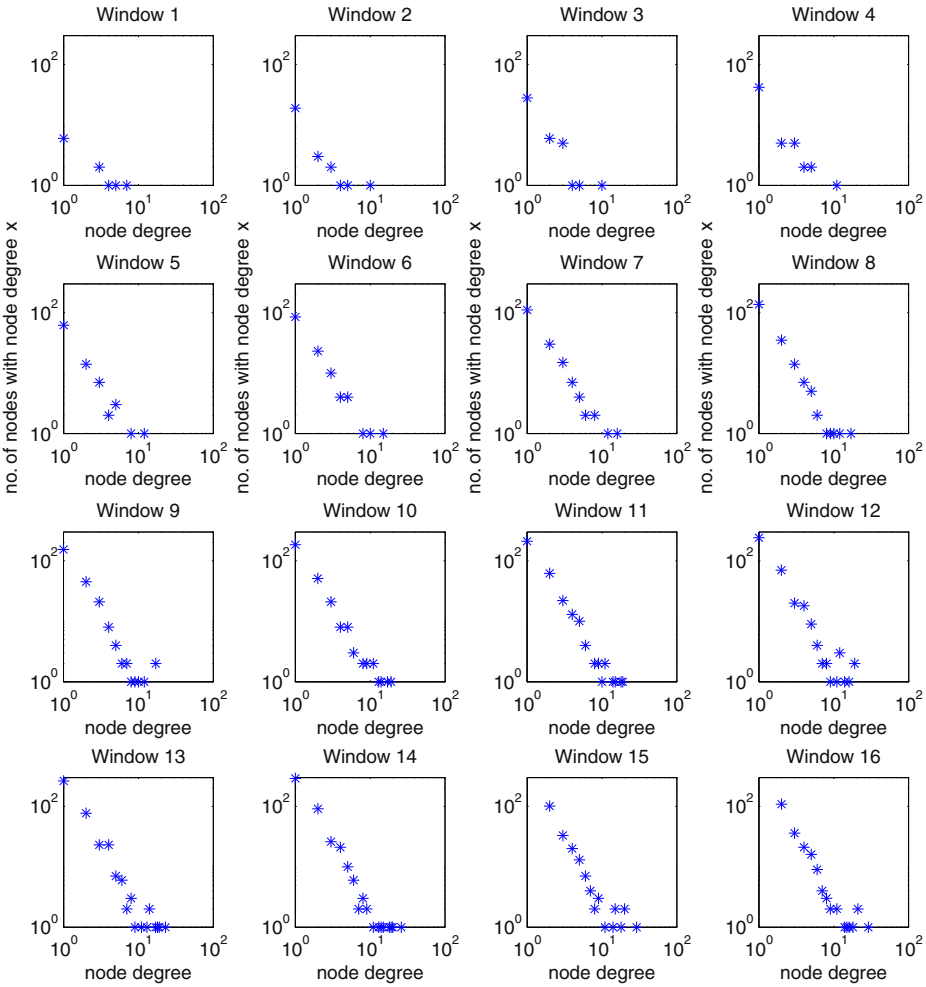

Figure 13 Degree distribution for merged networks for growing windows.

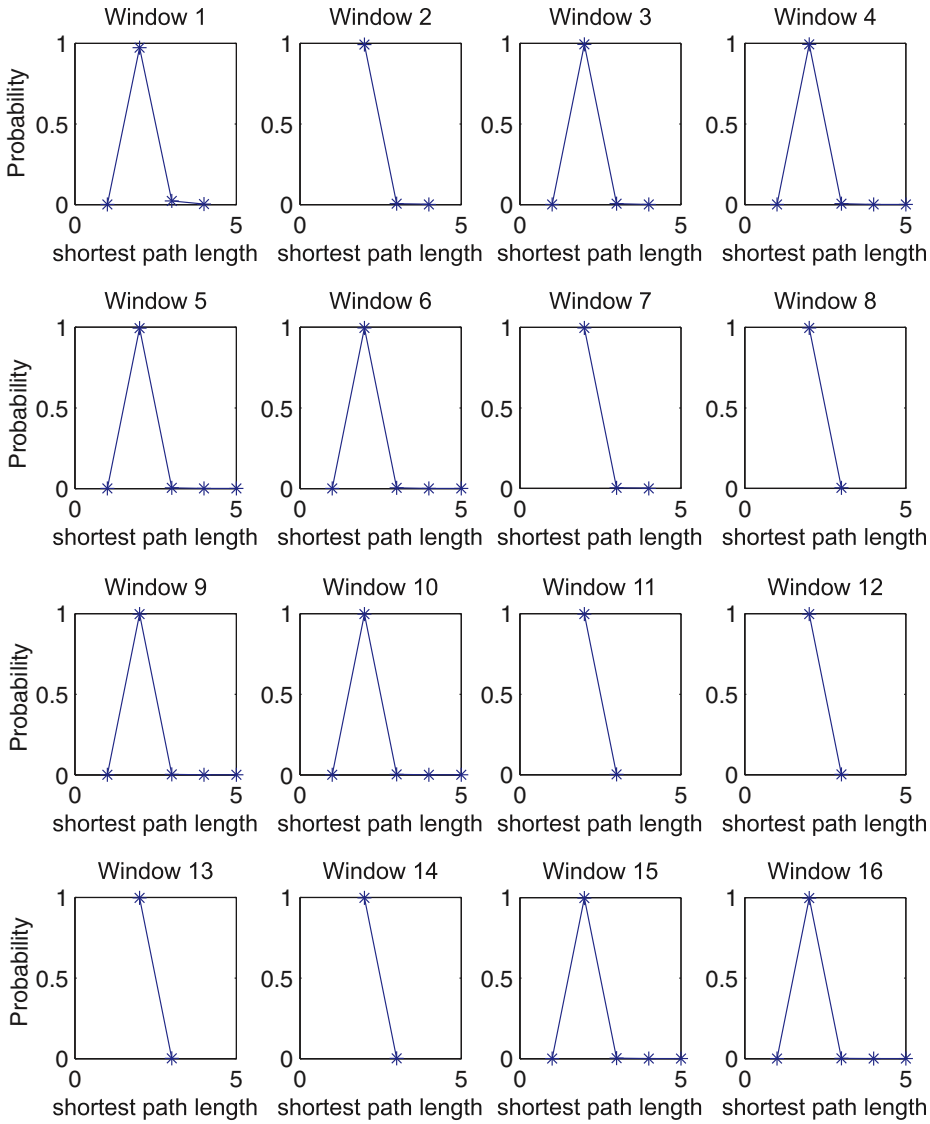

Figure 14 Shortest paths length distribution for contact network for growing windows.

small values—2 for contact network, 5 for forum network and 4 for merged networks and, which confirms the small-world phenomenon.

4.3.4 Dynamics of clustering coefficient

The dynamics of clustering coefficient (CC) is presented in Table 4 and further in Figure 17. Both Figure 17 and Table 4 include information about the mean clustering coefficient for a given network in each time window (created using growing window approach). The clustering coefficient grows in time for all networks but it is relatively

small; for the social networks it is expected to be higher [20]. The smallest values of the clustering coefficient are for the contact-forum network where it is at the level of random networks. In consequence the contact-forum network does not feature the phenomena where "fiend of my friend is my friend".

4.3.5 Dynamics of local topology

Another element that has been investigated is the local topology of the extracted networks.

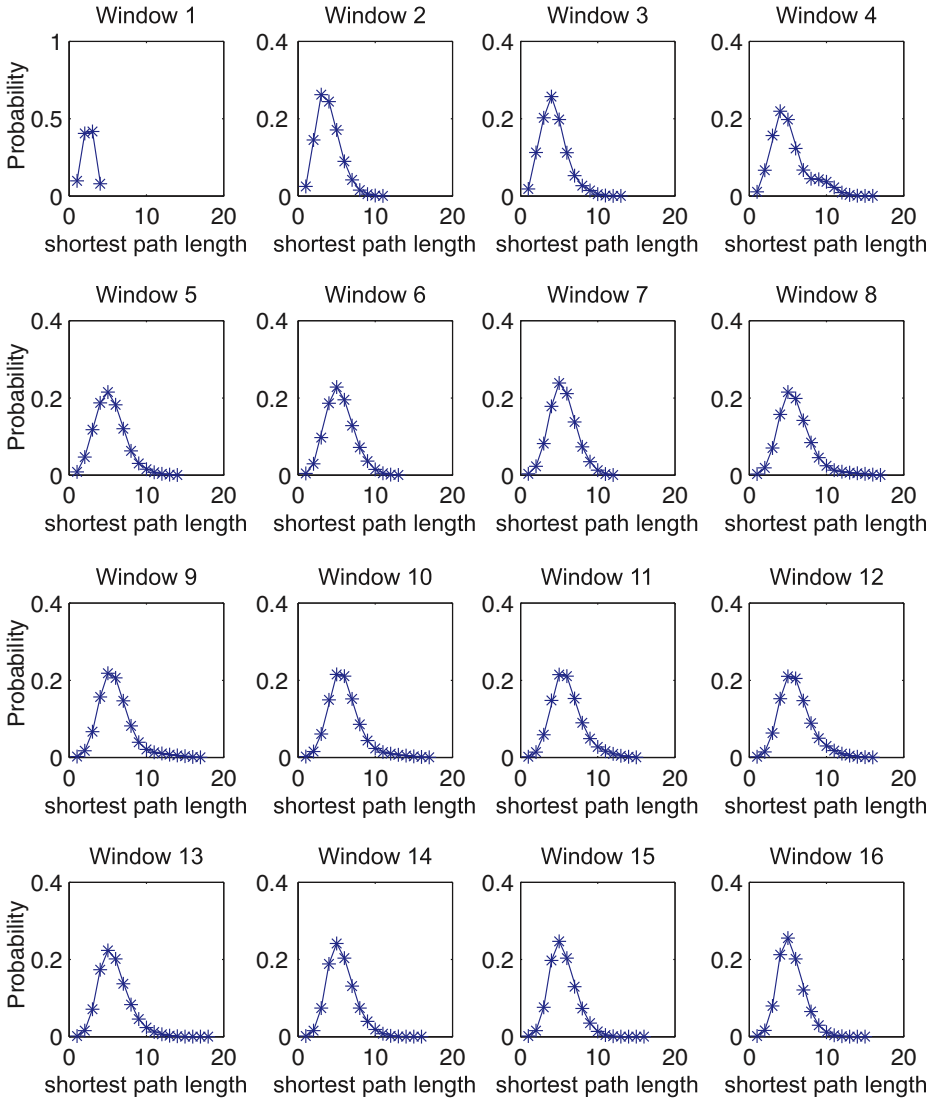

Figure 15 Shorthest paths length distribution for forum network for growing windows.

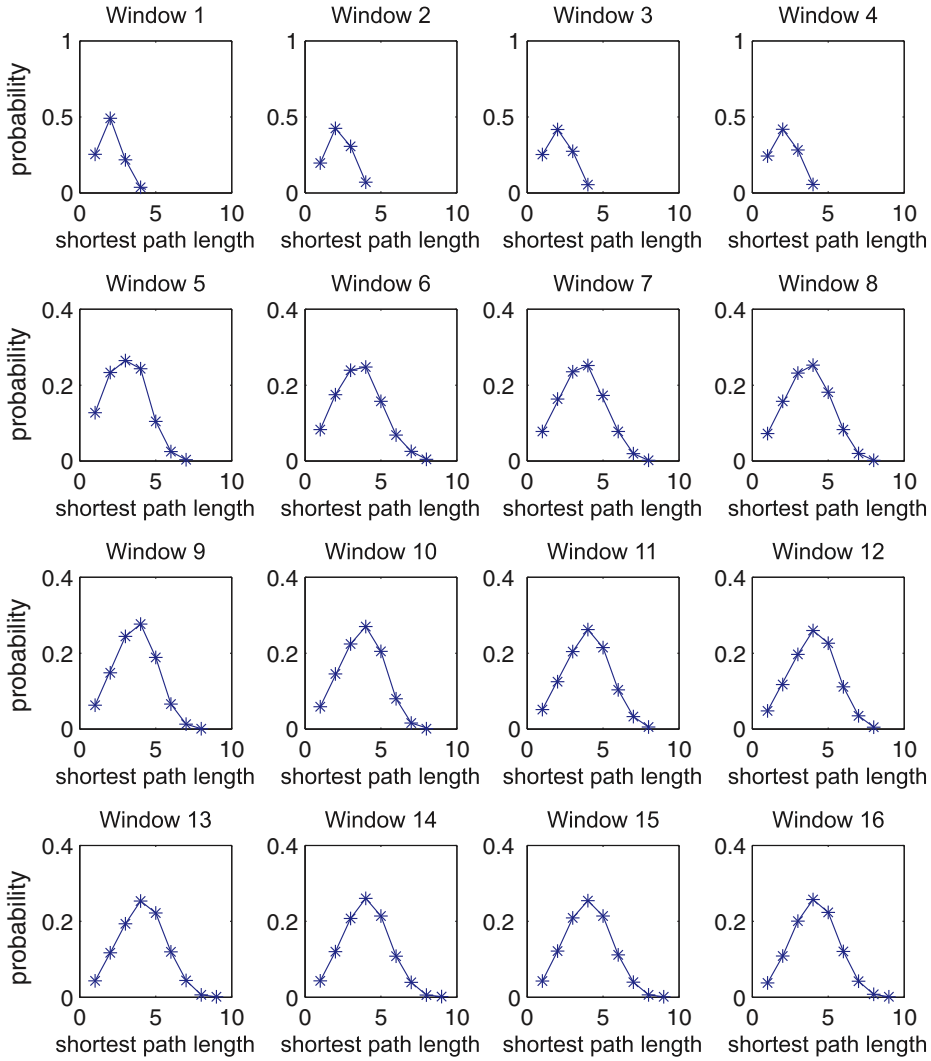

Figure 16 Shorthest paths length distribution for merged networks for growing windows.

Triad analysis was performed for all three networks that were split using separate and growing time windows. The results are presented in Figure 18. Each line denotes one time window; the lines at the bottom of each chart reflect Window 1 and when we move up along the *y* axes the successive lines represent the successive time windows. In all charts the number of triads grows with the growth of the network in time.

It can be seen that when the dataset is split into separate windows the number of specific triads grows systematically in the case of forum and contact network (Figure 18a and c). The growth can be also noticed in the case of contact-forum network but it is not regular (Figure 18e). When the network is analysed using the concept of the growing window, the number of specific triads grows linearly with time (Figure 18b, d, f).

Table 4 Mean clustering coefficient for different networks in specific time windows.

Window no.	Forum net	Contact net	Forum-contact net
1	0.0022	0.0095	$9.55 \cdot 10^{-4}$
2	0.0015	0.0113	$2.17 \cdot 10^{-4}$
3	0.0018	0.0198	$2.03 \cdot 10^{-4}$
4	0.0020	0.0220	$1.60 \cdot 10^{-4}$
5	0.0021	0.0212	$2.12 \cdot 10^{-4}$
6	0.0028	0.0227	$2.64 \cdot 10^{-4}$
7	0.0031	0.0251	$2.90 \cdot 10^{-4}$
8	0.0035	0.0256	$2.19 \cdot 10^{-4}$
9	0.0036	0.0249	$3.58 \cdot 10^{-4}$
10	0.0038	0.0249	$3.76 \cdot 10^{-4}$
11	0.0045	0.0268	$4.20 \cdot 10^{-4}$
12	0.0051	0.0282	$4.37 \cdot 10^{-4}$
13	0.0060	0.0295	$4.83 \cdot 10^{-4}$
14	0.0069	0.0315	$5.21 \cdot 10^{-4}$
15	0.0081	0.0322	$5.54 \cdot 10^{-4}$
16	0.0100	0.0336	$5.51 \cdot 10^{-4}$

In the system there are no triads of type 1, i.e. there are no isolated nodes. The interesting outcome of the analysis is that triad number 3, 5, and 7 in all networks are more frequent than other triads, which is a property stable over time. The element

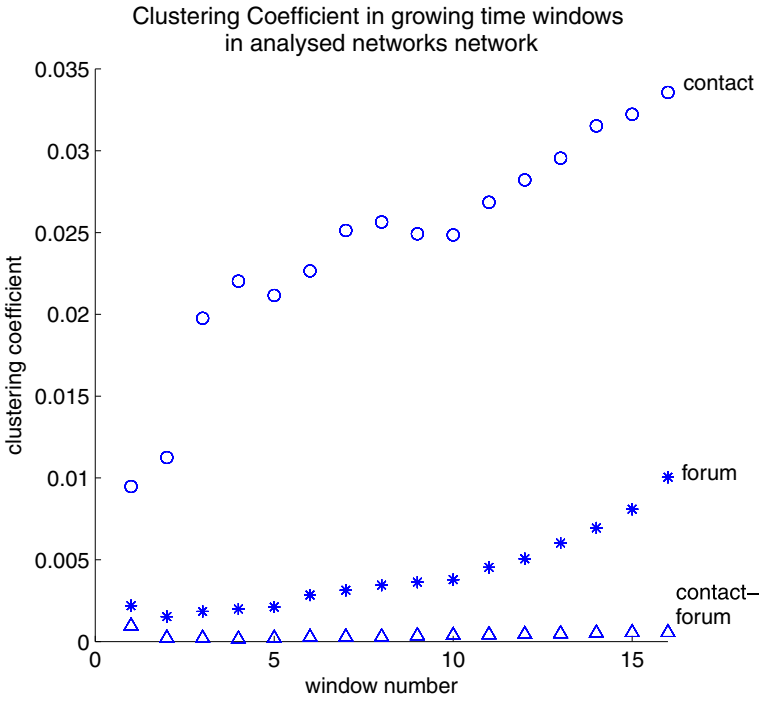

Figure 17 Dynamics of the mean clustering coefficient in forum, contact and merged networks for growing windows.

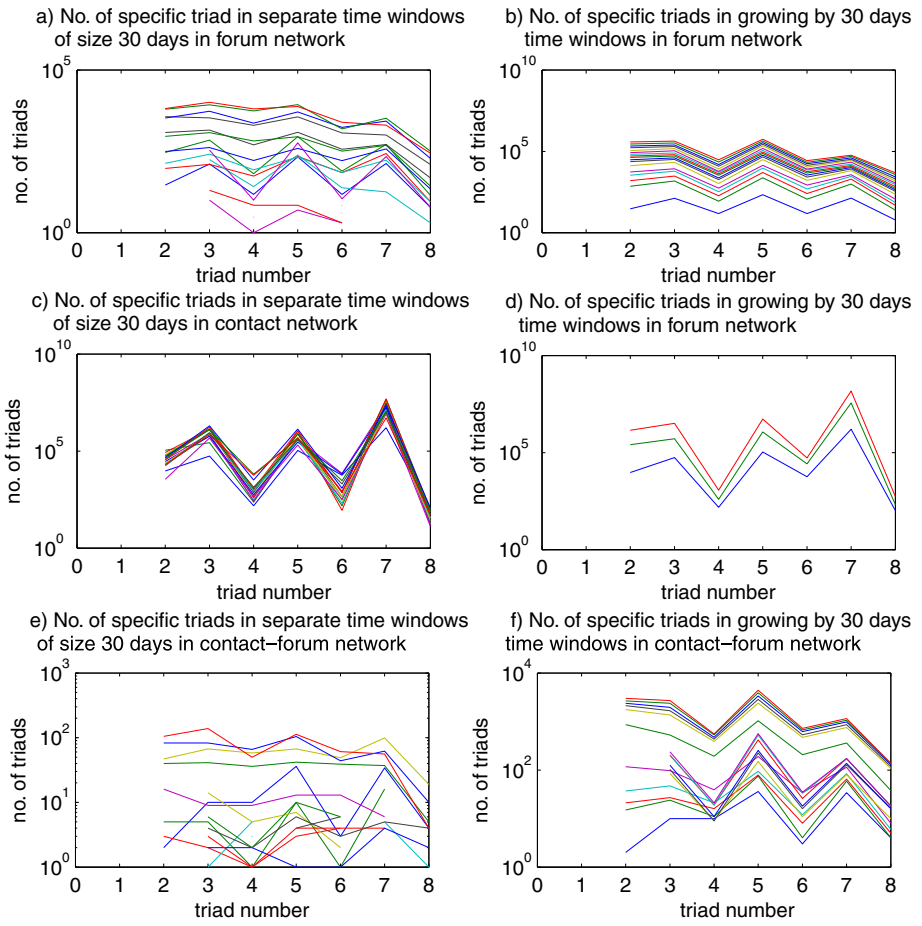

Figure 18 Triads for forum/contact/merged networks for separate and growing windows.

that is surprising is that triads number 4, 6, and 7 are more frequent than triad 8. In social networks triad 8, where all three nodes are connected, is perceived as very common and it is referred to as the "friend of my friend is also my friend" phenomenon. For sparse networks the occurrence of the triad 8 is very low (close or equal to zero in large random networks) and its appearance in our dataset clearly suggests above-average clustering met in most social networks. On the other hand, triads number 4, 6, and 7 is usually not present in social networks. Although triads number 4, 6, and 7 are structurally equivalent their frequencies differ. It suggests that the network topology is not regular and for user *a*, who is an element of many triads of type 7, his neighbours are not highly connected. This results in a local star topology and indicates that the user id 87 was assigned *a* position during triad counting. Thus the number of triads suggests that there are differences in the roles the users play in the local network structure. In the case of triad count with node identification the structurally equivalent triads should appear in similar numbers if the user roles (position in triads) are equally distributed.

5 Conclusions

The presented study aimed at investigating the dynamics of social network evolution. The experiments, using the real-world multirelational social network revealed that from the global perspective the network evolves following the scale-free paradigm. Although at the beginning the growth is random-like, for each analysed network a point in time can be identified after which the networks start evolving as a scale-free network. This point can be seen as phase transition for these networks.

The clustering coefficient for all networks is very low at the beginning but it grows with time, which is also visible in the growing number of fully connected triads. Nevertheless, it is not as high as in the case of regular social networks. On the other hand the local analysis of the networks using triad dynamics showed that the "friend of a friend" phenomenon is absent. This indicates that the network is not "social" in the traditional sense, i.e. with high clustering and small-world properties. This can be a result of the purpose for which the system has been created. Although it provides services which aim to enhance the collaboration (i.e. sharing, commenting, posting, responding to others opinions), the users are mainly looking for an advise from the specialist (user id 87) or read what others have published. An indicator of this is the fact that all analysed network are very sparse. Nevertheless, the short length of the shortest paths in all networks shows that they feature small-world phenomenon in terms of reachability in social networks, i.e. the process of reaching any node in the network is quick and possible to realise in few steps. Taking all of this into account, although the analysed networks feature the small-world effect, they are not clustered and these two characteristics are typical for random networks. The similar results were obtained while analysing vimeo system where people can follow other users, tag and add videos to their likes. The results also showed rather low clustering coefficient, lack of reciprocity and low networks density [18].

The analysis of actual dynamics of formation and growth of complex networks gives us opportunity to model these networks in a more precise way and this in turn enables us to describe and simulate such phenomena as spread of information and ideas, disease spread, trust and reputation in networks, etc. Modelling online social networks and investigation into their dynamics can be used to address such problems of Wisdom Web as personalisation, social and psychological context, coordination or semantics [29]. The investigation into the dynamics of people interactions can help to progress the work on such challenges as users' needs, common knowledge, or things' relations [29].

One may also notice that in the case of a typical web portal the results of social network analysis, especially when performed starting from the creation of the network, may be different from the huge and popular portals like Facebook, Flickr or Youtube. Moreover, considering more than one relationship type, different networks in terms of scale and structure are obtained, which calls for heuristics for merging networks composed of relations built on different bases.

A final conclusion of this study is that networks should be always investigated using various metrics and measures as only a combination of them provides the full picture and sufficiently detailed view on network characteristics and dynamics.

Acknowledgements The research presented in this paper has been partially supported by the European Union within the European Regional Development Fund program no. POIG.01.03.01-00-008/08.

References

1. Barabasi, A.L.: Linked: How Everything is Connected to Everything else and What it Means. Plume (2003)
2. Barrat, A., Barthelemy, M., Vespignani, A.: Dynamical Processes on Complex Networks. Cambridge University Press (2008)
3. Bisgin, H., Agarwal, N., Xu, X.: A study of homophily on social media. World Wide Web **15**(2), 213–232 (2012)
4. Bollobas, B.: Random Graphs. Academic, London (1985)
5. Braha, D., Bar-Yam, Y.: From centrality to temporary fame: dynamic centrality in complex networks. Complexity **12**, 59–63 (2006)
6. Breiger, R.: The analysis of social networks. In: Handbook of Data Analysis, pp. 505–526. SAGE Publications (2004)
7. Bringmann, B., Berlingero, M., Bonch, F., Gionis, A.: Learning and predicting the evolution of social networks. IEEE Intell. Syst. **25**(4), 26–35 (2010)
8. Carrington, P., Scott, J., Wasserman, S.: Social networks. In: Models and Methods in Social Network Analysis. Cambridge University Press, Cambridge (2005)
9. Fredericks, K., Durlan, M.: The historical evolution and basic concepts of social network analysis. New Dir. Eval. **2005**(107), 15–23 (2006)
10. Hill, S., Braha, D.: Dynamic model of time-dependent complex networks. Phys. Rev. E **82**, 046105 (2010). arXiv:0901.4407v2
11. Holland, J.: Hidden Order: How Adaptation Builds Complexity. Basic Books (1996)
12. Kazienko, P., Musial, K., Kajdanowicz, T.: Multidimensional social network and its application to the social recommender system. IEEE Trans. Syst. Man Cybern., Part A, Syst. Humans **41**(4), 746–759 (2011)
13. Kazienko, P., Musial, K., Brodka, P., Skibicki, K.: Analysis of neighbourhoods in multi-layered social networks. J. Comput. Intell. Syst. **5**(3), 582–596 (2012). doi:10.1080/18756891.2012.696922
14. Kumar, R., Novak, J., Tomkins, A.: Microscopic evolution of social network. In: The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press (2006)
15. Lescovec, J., Backstrom, L., Kumar, R., Tomkins, A.: Microscopic evolution of social networks. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) (2008)
16. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. J. Am. Soc. Inf. Sci. Technol. **58**(7), 1019–1031 (2007)
17. Musial, K., Kazienko, P.: Social networks on the internet. World Wide Web J. 1–42 (2012). doi:10.1007/s11280-011-0155-z, online first
18. Musial, K., Sastry, N.: Social media—are they underpinned by social or interest-based interactions? In: 4th Annual Workshop on Simplifying Complex Networks for Practitioners (SIMPLEX2012) Co-Located with World Wide Web Conference (WWW2012) (2012)
19. Newman, M.E.J.: Assortative mixing in networks. Phys. Rev. E **89**(20), 208701 (2002)
20. Newman, M., Park, J.: Why social networks are different from other types of networks. Phys. Rev. E **68**(3), 036122 (2003)
21. Shen, H.T., Hua, X.S., Luo, J., Oria, V.: Guest editorial: content, concept and context mining in social media. World Wide Web **15**(2), 115–116 (2012)
22. Time.com: One minute on Facebook—person of the year 2010—Time (2011)
23. Wasserman, S., Faust, K.: Social Network Analysis Methods and Applications. Cambridge University Press, New York (1994)
24. Watts, D.: Small Worlds Dynamic of Networks between Order and Randomness. Princeton University Press (2002)
25. Watts, D., Strogatz, S.: Collective dynamics of small-world networks. Nature **393**(6684), 440–444 (1998)

26. Yao, J., Cui, B., Huang, Y., Zhou, Y.: Bursty event detection from collaborative tags. World Wide Web **15**(2), 171–195 (2012)
27. Zhong, N., Liu, J., Yao, Y.Y.: In search of the Wisdom Web. IEEE Computer **35**(11), 27–31 (2002)
28. Zhong, N., Liu, J., Yao, Y.Y.: Web Intelligence (WI). In: The Encyclopedia of Computer Science and Engineering, vol. 5, pp. 3062–3072. Wiley (2009)
29. Zhong, N., Ma, J.H., Huang, R.H., Liu, J.M., Yao, Y.Y., Zhang, Y.X., Chen, J.H.: Research challenges and perspectives on Wisdom Web of Things (W2T). J. Supercomputing (Springer) (2010). doi:10.1007/s11227-010-0518-8